Journal of the Royal Society of Western Australia, 92: 431–435, 2009

Darwin's Tree of Life and challenges from modern genetics

K R Thiele

Western Australian Herbarium
Department of Environment & Conservation,
Science Division,
Locked Bag 104 Bentley DC, WA 6983

Manuscript received January 2010; accepted February 2010

Abstract

The tree of life has been a central organising principle in biology for centuries. Darwin provided a mechanism – descent with modification – through which a tree of life could be explained and understood. Since his time a central project of biology has been the elucidation of the structure of the tree (phylogenetics). Recently, following the success of molecular data in allowing very wide and deep comparisons of taxa from throughout life, challenges to the metaphor of life as a tree have emerged. The challenges arise from observations that not all inheritance of genes, and hence of characteristics, is through descent; some alleles in a given taxon may be derived by lateral transfer from another organism, sometimes in a very different domain of life. This reticulate inheritance, if common and widespread, would mean that life would be better viewed as a net than a tree. This paper provides an empirical argument why, at least for the multicellular domains of life, rejection of a tree of life is premature and tree-like inheritance through descent with modification predominates over reticulate inheritance through lateral gene transfer.

Keywords: Darwin, phylogenetics, Tree of Life, genetics, lateral gene transfer

Introduction

A recent series of papers (Doolittle 1999; Dagan & Martin 2006), popularized in a cover article in *New Scientist* during the Darwin bicentenary (Lawton 2009), have made a strong claim: that Darwin's idea of a 'tree of life', an overarching concept in evolutionary biology since publication of *On The Origin of Species*, is wrong.

The concept of a single tree of life was central to Darwin's thinking and to the development of his theory of evolution by natural selection (Padian 2008). A core project of evolutionary biology for the last 150 years has been to elucidate details of the tree. However, the *New Scientist* article claimed that "today the project lies in tatters, torn to pieces by an onslaught of negative evidence", that "many biologists now argue that the tree concept is obsolete and needs to be discarded" and that "the tree of life ... has turned out to be a figment of our imagination".

This paper reviews the conceptual history and antecedents of the tree of life, discusses the challenge to the concept provided by new evidence, and argues that, at least in the multicellular domains of life, the pattern of variation observed in the real world provides strong evidence that life does indeed have a tree-like structure and that Darwin's tree of life continues to be a useful concept.

Darwin's Tree of Life and its antecedents

The first evidence that Darwin viewed life as tree-like comes from a sketch on p.36 of his Notebook B (*c.* July 1837, one year after the return of the *Beagle* to England; see van Wyhe, 2002–9). The sketch shows a branching diagram rooted at the base, each node having two or more branches arising from it. Some terminals in the tree bear cross-bars (representing extant taxa) while others lack them (representing extinct taxa). The accompanying notes show that Darwin was considering concepts of relatedness (some taxa being more closely related than others) and extinction (particularly the need to keep the numbers of species constant from one generation to another).

Three principal antecedents to Darwin's tree are likely to have influenced his thinking. The first is the *scala naturae* (often called the *Great Chain of Being*), a mediaeval concept of the universe in which all nature is arranged in a natural and immutable hierarchical order, with the inanimate world at base and God at the apex. In intermediate positions (from the base, in order) are vegetable forms, animals (sometimes in order from least to most useful), commoners, nobles, angels and apostles. Many representations of the *scala naturae* are found in mediaeval literature, and the concept would undoubtedly have been familiar to Darwin. While the *scala naturae* is not strictly tree-like, it is rigidly hierarchical and accommodates all natural forms in a single structure, properties of Darwin's tree also.

The second antecedent is the classical system of classification (categorization) developed by Greek philosophers, particularly Plato and Aristotle, and applied to many natural and conceptual systems from the mediaeval period onwards. In classical categorization, classes are definable, exclusive and exhaustive: that is, any class should belong unambiguously to one and only one higher-order class, and any class should contain all subclasses that have the same defining properties.

Beginning with Aristotle, the natural world was seen by many natural historians as a model system for classical categorization. Increasingly sophisticated systems were developed to accommodate knowledge, in the biological realm culminating in Linnaeus' *Systema Naturae*. While generally conceptualised during this period as a "boxes within boxes" system, a classical classification may also be represented as a tree, with the most inclusive category as the root and any node having only one parent and at least one child. Again, Darwin was clearly very familiar with classical categorization, discussing it at length in the *Origin*, and Darwin's tree of life shares the same properties.

The third antecedent, again one with which Darwin would have been very familiar, is the biological key, developed from "tables" of relationships around the middle of the Seventeenth Century (see Nelson & Platnick 1981 for a discussion of the history of keys and other biological tree structures). Keys are a special class of dichotomizing classification, and are often represented in a tree-like form with a root and a branching, tree-like structure. Biological keys are closely related to classical categorizations, and the two are often more or less interchangeable.

With these antecedents, Darwin would have been very familiar with the representation of knowledge as a tree and with the use of trees to represent biological systems and taxonomic relationships. The tree first sketched in Darwin's notebook in 1837 became the sole figure in the *Origin*, an indication of the importance he placed in the concept of a tree of life.

In Darwin's tree and most of its antecedents, every node (except the root) has exactly one ancestor node and two or more descendent nodes, and for every node (taxon) there is a single line of descent from the root. This presupposes four main conceptual properties of life. Firstly, the fact that every taxon is connected to every other precludes the pre-Darwinian concept of spontaneous generation, in which new living organisms arise *de novo* from non-living matter. Secondly, the fact that the tree is rooted presupposes that evolution is unidirectional; once extinct, a taxon cannot be recreated. Thirdly, the fact that every taxon has only one direct ancestor presupposes that a given taxon cannot arise twice independently. Lastly, the fact that every taxon has a single line of descent presupposes that all inheritance is by descent (with modification).

Calculating Darwin's tree

Darwin's critical concept of descent with modification, one of the two core ideas of the *Origin* (the other was natural selection), has been one of the most productive ideas in biology. It provided a theoretical basis for classical biological categorization and, by providing an argument for the existence of a natural classification (one which reflects the pattern of evolutionary descent) greatly contributed to the development of post-Linnaean classification systems. Understanding evolutionary patterns in turn has provided an underpinning framework for virtually the whole of biology.

For close to a century after publication of the *Origin*, the tree of life acted more as a guiding principle than as a methodological tool. Many alternative systems of classification were erected, each claiming to be natural, but there was no method for rigorously testing classifications for goodness of fit with the actual pattern of evolution or with the observed pattern of distribution of characters, nor could alternative classifications be compared on any grounds other than utility and preference. Trees were frequently drawn, most notably those of Haeckel (Haeckel 1866), but more as artistic expressions of a set of loose ideas of relationships than as rigorous, testable hypotheses.

This situation changed dramatically in the 1950s with the development of cladistic (phylogenetic) methodologies by Hennig (1950; 1966). Hennig realized that descent with modification provided a powerful tool for constructing hypothetical trees of life (phylogenies) and for testing these hypotheses against independent data sets. The breakthrough came in the recognition that some shared characteristics – those shared because they are novelties or modifications within the evolutionary history of the group concerned – are informative of relatedness, while other shared characters – those shared because they are unmodified from the ancestral state – are not informative of relatedness at the level of interest. Following Hennig, mathematical algorithms were developed to derive trees from sets of observations of character states in taxa (see Cavalli-Sforza & Edwards 1967; Nei 1996 for reviews).

The conceptual basis for phylogenetic analysis is simple. If evolution by descent with modification has occurred, then character states that have become modified during the evolution of a group should be distributed within the terminal taxa of the group in a way that encodes information about the underlying pattern of evolution (Figure 1). Phylogenetic analysis seeks to uncover a pattern of branching that best explains the observed pattern of distribution of character states in these terminals. This is done using a nominated criterion of goodness of fit, such as parsimony, likelihood or

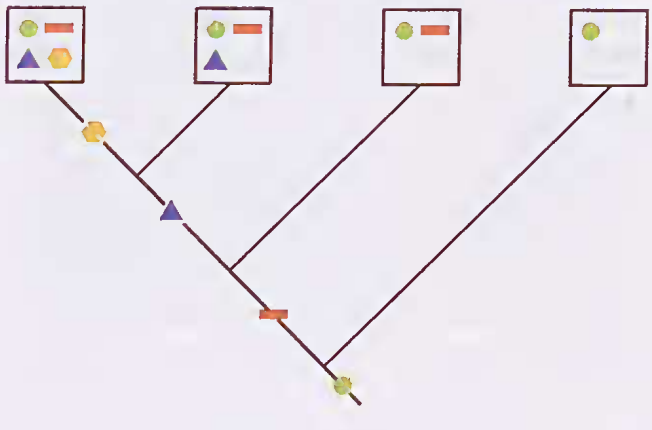

Figure 1. A simple, imaginary phylogeny of four organisms. Boxes represent extant terminal taxa; lines represent the lines of descent of the taxa; symbols represent modifications of character states, showing the pattern of descent and the distribution of states in the terminal taxa. Phylogenetic analyses are able to reconstruct the pattern of branching by analysing the distribution of states in the terminal taxa.

bayesian probability. Of all the possible ways of arranging the taxa of interest into a tree, some trees provide a good explanation (under the nominated criterion) while others provide a poor one: the analysis methods search the set of possible trees for the best. An implicit assumption is that trees derived from such algorithms are candidates for parts of the tree of life.

All methods for constructing phylogenetic trees allow that the best tree found may not explain all aspects of the observed distribution of character states. This is necessary because noise may creep into the phylogenetic signal through several processes. The two main sources of noise are errors of interpretation of character states (errors of homology), and evolutionary processes other than direct descent with modification, such as lateral transfer of character states derived in one part of the tree directly to another part of the tree through hybridization and direct gene transfer.

Noise is identified, and accounted for, by assessments of congruence. A lack of congruence between two or more characters indicates that at least one of the characters is providing noise rather than signal to the analysis (Figure 2). Noise is filtered from the analysis by preferring trees that maximize congruence among characters.

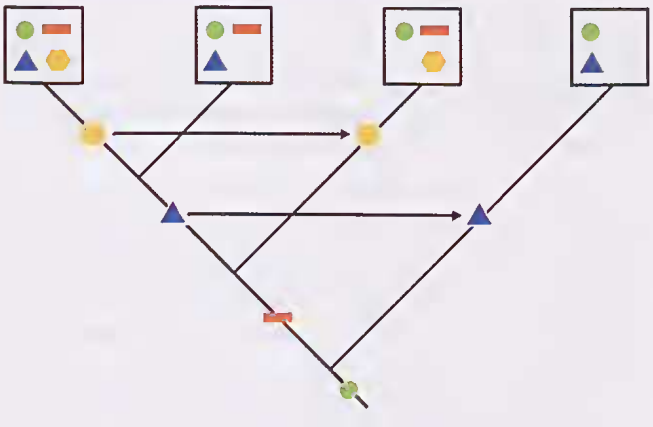

Figure 2. The phylogeny of Figure 1 with noise added to the phylogenetic signal through lateral gene transfer. Phylogenetic analyses are usually still able to reconstruct the pattern of branching by analysing the distribution of states in the terminal taxa, taking into account congruence between characters, as long as the signal to noise ratio is not too severely compromised.

Two developments in the latter half of the twentieth century had a profound effect on phylogenetics, taking it into realms Darwin would have found unimaginable. The first was the development of the methods outlined above. The second was advances in gene technology, particularly in rapid and inexpensive DNA and RNA sequencing methods. These two factors – the ability to generate enormous amounts of potentially informative phylogenetic data and the ability of analyse the data rigorously – led to an explosion of interest in phylogenetics and the tree of life. The same factors also led to challenges to the validity of the entire conceptualization of life as a tree.

Challenging Darwin's tree

A major strength of genetic data in phylogenetic reconstruction is that all living organisms share many genes. This provides a depth and power to phylogenetic analysis that was impossible to gain from earlier morphological studies. Using appropriate molecular sequences, phylogenetic analyses can assess relationships between, for example, species of bacteria, algae, flowering plants and whales. The diversity of form across this set of organisms makes morphological comparison between them meaningless; the conservatism of their sequences, conversely, makes genetic comparisons richly meaningful.

As more and more sequences were accrued and such phylogenetically deep comparisons became commonplace from the early 1990s, incongruous examples of sequence relationships began to accumulate (see Doolittle 1999 for a review). Specific gene sequences that were postulated to have evolved in the Archaea were found in Eubacteria, and vice versa. It is now widely accepted that lateral gene transfer in and between eubacteria and archaebacteria is widespread (Jain *et al.* 1999; Dagan *et al.* 2008), and a number of mechanisms by which it occurs have been elucidated, including transfer of genes via virus and bacteriophage vectors and direct uptake of DNA fragments from the environment (Martin 1999; Thomas & Nielsen 2005). However, its frequency is not well established. Estimates of the proportion of prokaryotic genes that have been transferred laterally at least once vary from 2–60% (Ge *et al.* 2005; Lerat *et al.* 2005).

The role of lateral gene transfer in eukaryotes, particularly multicellular eukaryotes, is less clear. Sequences believed to have evolved in bacteria after the symbiosis event that gave rise to eukaryotes have been found in eukaryotic taxa (Keeling & Palmer 2008; Ros & Hurst 2009), and sequences believed to arise in one branch of eukaryotic life have been found in others (*e.g.* Lisch 2008; Alsmark *et al.* 2009). Cases of transfer of specific sequences into widely unrelated groups (such as between bacteria and mammals) presumably occurred deep in evolutionary history. Other mechanisms for transfer that have been postulated in eukaryotes include hybridisation, which appears to be common in some groups such as plants (although again its frequency is highly uncertain; see Ellstrand 1996 for a review). Transfer through hybridisation events may potentially involve large stretches of DNA including many genes.

Lateral gene transfer directly questions the validity of the tree of life. If evolution of organisms is principally driven by descent with modification, then a tree is an appropriate metaphor and representation. If, however, most evolution is driven by lateral transfers of genes then, as some authors have claimed, life is a web (Hilario & Gogarten 1993) and viewing it as a tree is counterproductive. This view has been claimed as a paradigm shift in biology by some (O'Malley & Boucher 2005).

Testing Darwin's tree

Rather than arguing from first principles whether trees or nets are the best representations for the phylogenetic structure of life, it would be useful to have an empirical test of the goodness of fit of different representations. This in turn would provide some indication of the significance (as against the mere existence) of hybridization and lateral gene transfer in the evolution of any group of organisms. In fact, such an assessment is possible, and indeed is inherent in the structure of phylogenetic analysis.

Consider an imaginary world in which lateral gene transfer greatly predominates over descent with modification, and is random. The pattern of evolution would be a closely reticulating net, with genes (and hence their phenotypic expression as characteristics of organisms) jumping widely and randomly between different strands of the net during evolutionary time. A result of such frequent transfer would be that alleles and their characteristics will be distributed randomly across extant taxa. In such a world, phylogenetic analysis using algorithms that seek to identify a tree-like pattern would be characterized by gross incongruence. Two or more analyses based on independent subsamples of characters or molecular sequences would be very likely to give different (incongruent) tree topologies.

Compare this with a second imaginary world in which descent with modification greatly predominates over lateral gene transfer. The opposite outcome would be expected – alleles and character states would be highly patterned in extant taxa, and two or more phylogenetic analyses based on independent subsamples of data would usually give similar (congruent) topologies.

How does the real world compare with these two imaginary worlds? In general amongst plants and multicellular animals there is a striking degree of congruence between independent phylogenetic analyses of the same group of taxa, as indeed there is between modern phylogenetic analyses and classifications based on traditional, pre-phylogenetic inferences. For example, in a recent change from a 90-year old classification (based on Engler & Prantl 1905) to a modern classification (APG III 1990) at the Western Australian Herbarium, there were 53 cases spread over 226 families in which genera or groups of genera were reallocated to new families.

On one hand, this result could be regarded as indicating a substantial problem of lack of congruence between the traditional and modern classification (involving 23% of families). However, almost all of the cases of incongruence can be explained by factors other than lateral gene transfer, particularly the failure of traditional classifications to adequately reflect relationships by descent (monophyly), difficulties in accurately interpreting homologies, and sampling artefacts. Thus, there is a >>80% congruence between old and new classifications. Similarly, the APG III system itself is based on the striking congruence between independent data sets, and is regarded as approaching stability in most areas (APG III 1990).

Even in the bacterial domains, attempts to extract congruent tree-like patterns from amongst the noise of random lateral transfers appear successful (Puigbò *et al.* 2009). The world observed through phylogenetic analysis appears to be a world in which lateral gene transfer is not so rife as to destroy the tree of life. Of course, if lateral gene transfer is non-random, instead involving coordinated jumps of large parts of genomes, then the observed phylogenetic congruence in the real world cannot be regarded as evidence against ubiquitous lateral gene transfer. However, most proposed mechanisms for transfer (infections by viruses, hybridization and direct uptake of DNA from the environment) are essentially random events, and empirical evidence suggests that, while they undoubtedly occur, they do not significantly challenge the validity of Darwin's tree.

Conclusions

In multicellular life, congruence between trees obtained by phylogenetic analysis of independent data sets is commonplace. Indeed, cases of incongruence are intensely interesting and are the exceptions that prove the rule. This provides empirical evidence that random lateral gene transfer is not rife, although it undoubtedly exists. This in turn suggests that a tree-form representation of life – Darwin's tree – is an appropriate metaphor and framework principle. It would be premature to reject it at this stage.

The demonstrated existence of some lateral gene transfer shows that some reticulations do occur in the tree of life. Such reticulations may be common in some groups, particularly among prokaryotes, but it should not be assumed that they break the tree metaphor until it can be demonstrated in those groups that phylogenetic congruence is rare rather than common.

Most methods of phylogenetic analysis assume a tree-like structure and force a tree-like result. Work to find analysis methods that can handle reticulations continues (see *e.g.* Linder & Rieseberg 2004; Reeves & Richards 2007). However, if phylogenetically meaningful network algorithms can be developed we will have a situation in phylogenetics akin to that in astronomy. Radio astronomers and visible-light astronomers do not argue as to which method provides the best view of the universe. Rather, their telescopes are regarded as complementary. Similarly, the evolution of life appears to have tree-like and network-like aspects, and the use of "telescopes" that recover tree-like signal and net-like signal are complementary rather than competitive approaches.

References

Alsmark U C, Sicheritz-Ponten T, Foster P G, Hirt R P, Embley T M 2009 Horizontal Gene Transfer in Eukaryotic Parasites: A Case Study of *Entamoeba histolytica* and *Trichomonas vaginalis*. Methods in Molecular Biology 532: 489–500.

APG III 1990 An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Botanical Journal of the Linnean Society 161: 105–121.

Cavalli-Sforza L L & Edwards A W F 1967 Phylogenetic analysis. Models and estimation procedures. American Journal of Human Genetics 19: 233–257.

Dagan T & Martin W 2006 The tree of one percent. Genome Biology 7: 118.

Dagan T, Artzy-Randrup Y, Martin W 2008 Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proceedings of the National Academy of Science USA 105: 10039–10044.

Doolittle W Ford 1999 Phylogenetic classification and the Universal Tree. Science 284: 2124–2128.

Ellstrand N C, Whitkus R & Rieseberg L H 1996 Distribution of spontaneous plant hybrids. Proceedings of the National Academy of Science USA 93: 5090–5093.

Engler A & Prantl K (eds.) 1887–1915 Die Natürlichen Pflanzenfamilien. Engelmann, Leipzig.

Ge F, Wang L S & Kim J 2005 The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biology 3: e316.

Haeckel E 1866 Generelle Morphologie der Organismen: allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von C. Darwin reformirte Decendenz-Theorie. Remer: Berlin.

Hennig W 1950 Grundzüge einer Theorie der phylogenetischen Systematik. Deutscher Zentralverlag: Berlin

Hennig W 1966 Phylogenetic Systematics (tr. D Davis and R Zangerl), University of Illinois Press: Urbana

Hilario E & Gogarten J P 1993 Horizontal transfer of ATPase genes – the tree of life becomes a net of life. BioSystems 31: 111–119.

Jain R, Rivera M C & Lake J A 1999 Horizontal gene transfer among genomes: the complexity hypothesis. Proceedings of the National Academy of Science USA 96: 3801–3806.

Keeling P J & Palmer J D 2008 Horizontal gene transfer in eukaryotic evolution. Nature Reviews Genetics 9: 605–618.

Lawton G 2009 Uprooting Darwin's tree. New Scientist 2692: 34–39.

Lerat E, Daubin V, Ochman H & Moran N A 2005 Evolutionary origins of genomic repertoires in bacteria. PLoS Biology 3: e130.

Linder C R & Rieseberg LH 2004 Reconstructing patterns of reticulate evolution in plants. American Journal of Botany 91: 1700–1708.

Lisch D 2008 A new SPIN on horizontal transfer. Proceedings of the National Academy of Science USA 105: 16827–16828.

Martin W 1999 Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. BioEssays 21: 99–104.

Nei M 1996 Phylogenetic analysis in molecular evolutionary genetics. Annual review of genetics 30: 371–403.

Nelson G & Platnick N 1981 Systematics and Biogeography: Cladistics and Vicariance. Columbia University Press. New York

O'Malley M A, Boucher Y 2005 Paradigm change in evolutionary microbiology. Studies in the History and Philosophy of Biological & Biomedical Science 36: 183–208.

Padian K 2008 Darwin's enduring legacy. Nature 451: 632–634

Puigbò P, Wolf Y I & Koonin E V 2009 Search for a 'Tree of Life' in the thicket of the phylogenetic forest. Journal of Biology 8:59.

Reeves P A & Richards C M 2007 Distinguishing terminal monophyletic groups from reticulate taxa: performance of phenetic, tree-based, and network procedures. Systematic Biology 56: 302–320.

Ros V I & Hurst G D 2009 Lateral gene transfer between prokaryotes and multicellular eukaryotes: ongoing and significant? BMC Biology 7:20.

Thomas C M & Nielsen K M (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nature Reviews Microbiology 3: 711–721

van Wyhe J 2002–9 The Complete Work of Charles Darwin Online [http://darwin-online.org.uk/content/frameset?viewtype=side&itemID=CUL-DAR121.-&pageseq=38]